iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37439

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Epidemic Modelling of COVID-19

Mayanka Gupta[1], Mrs. Renuka Malge[2]
[1]*P.G Scholar,* [2]*Asst. Professor, Artificial Intelline and Machine Learning, Visvesvaraya Technological University For Post Graduate Studies, Bangalore, India*

Abstract: ***COVID-19 has had a disastrous impact on millions of lives all over the world.199,466,211 confirmed cases of COVID-19 and 4,244,541 deaths have been reported to WHO till 4th august. Analyzing the available data and predicting the pandemic trend is important since the situation can be controlled only when there is adequate preparation. Research using epidemiological models helps in analyzing different facets of COVID including infection, recovery and death rate. Predicting the daily increase of cases can help reduce the burden on health care workers and government by aiding them in planning the required resources in advance. Thus, in this project data driven epidemic modelling approach is used. COVID cases of 10 forthcoming days using three modelling techniques namely Polynomial Regression, Bayesian Ridge Regression and Support Vector Machine are predicted. The performance metric used to identify the best model are MSE and MAE. Polynomial Regression is found to have best performance followed by Bayesian ridge regression. Support Vector Machine has a poor performance.***
Keywords: Epidemic Modelling, COVID-19, Machine Learning, Polynomial Regression, Bayesian Ridge Regression, Support Vector Machine

I. INTRODUCTION

COVID-19 pandemic is a communicable respiratory syndrome brought about by a corona virus called SARS-CoV-2. This pandemic has greatly disrupted everything from the daily life of general public to the Global Economy. Analysis of the data provides insights on what are some good measures to reduce the spread of the disease. Forecasting the probable number of infected cases gives a rough estimate of the amount of preparations that need to be made for providing the best healthcare services to all those who are affected during this pandemic. In the analysis task, government and researchers need to work hand in hand. Using Machine Learning and Data Analytics to the available data allows the study of spread patterns in the population. In this paper we explore the different types of epidemic models. The research carried out so far using these various approaches. Finally, the data of COVID-19 is modeled using Polynomial Regression, Bayesian Ridge Regression and Support Vector Machine to forecast the COVID cases for forthcoming 10 days. The performance of these models is measured using MSE and MAE.

II. LITERATURE SURVEY

Epidemic is described as widespread occurrence of an infectious disease in a community at a particular span of time. When the epidemic spreads over multiple countries, continents or throughout the globe it is referred to as pandemic. The mathematical and data analytical models can project how these infectious diseases progress to show the likely outcome of an epidemic/pandemic. This in turn helps the government and public health body to identify the required interventions to deal with the consequences of the pandemic. Basic assumptions are made to model the collected data using mathematical approach to find parameters for various infectious diseases. These parameters are then used to calculate the effects of different steps, like mass vaccination, quarantine, lockdown etc. The model helps decide which step is useful and to what extent. The model can also predict future growth patterns of the epidemic. Epidemic models can be classified into three types:

A. Compartmental Models

In a compartmental model, people are assigned to different categories or compartments, each representing a specific stage of the epidemic. The switch from one group to another are mathematically expressed as derivatives. Thus the model is formulated using differential equations. SIR model is the most basic and widely used compartmental model. There are many adaptations of the SIR model. In SIS model births and deaths during the pandemic are included and there is no immunity upon recovery. In SIRS model immunity lasts only for a short period of time. In SEIS and SEIR model there is a latent period of the disease where the person is not infectious. In the MSIR model infants can be born with immunity.

In [1] the author uses neural network to find best parameters for the SIR model. The author considers S, I and R as variables and β and γ are considered the parameters. Utilizing the time dependent parameters they arrive at the conclusion that data driven model like Runge-Kutta method can be functionally equated to the compartmental model such as SIR for the COVID data. The modelling was done on data for South Korea.

In [2] the author uses SIRD model to predict death rate. The two variants used mortality rate of 0.3 and 1.0. Social distancing is also introduced to reflect behavioral and policy changes due to pandemic. Data of New York and other U.S states is considered for this study.

In [3] the author studies lock down targeted at the population using age as the criterion. Multi-risk SIR model is used for the research. This model is efficient to limit the interactions of high risk category with the remaining population with an approach alike the lock down, thereby reducing the death rate. It also boosts economic recovery of the population.

In [4] six endemic models have been considered for infectious diseases. They have used three variations of SIQS and SIQR model each. SIQR model with the quarantine adjusted variant for some parameters showed an unstable spiral in the established equilibrium.

In [5] based on isolation five different categories of population was studied. The SEIR model was modified to include the isolation of patients suffering from COVID and the new model was named SEIIsR model. The paper also emphasizes the importance of lock down to reduce spread of infection in particular in Bangladesh. For UK a similar model was developed that included some of the government policy implementations. Maximum spread of infection was predicted for three regions in Italy that were first to be affected by the pandemic.

In [6] SIQR model is studied. This model is similar to SIR model but includes one more category quarantine. This paper shows that a single person is capable of infecting 1.55 other people if not quarantined. It studies the population of Brazil and Italy. The spread of infection is predicted using SEIR model for heterogeneous population.

B. Data Driven Model

Data driven models use the available data for the entire population instead of sorting people into categories. They analyze and infer the pattern for the spread of epidemic. Various Machine learning algorithms can be used to model the scenario. The model behaves as a black box as the modeler is unaware of the parameters used in the prediction. These parameters are derived as the representation of the complex relationships among a broad set of inputs that are used to predict certain outputs. Various ML techniques that are used to predict and forecast future events are SVM, linear regression, logistic regression, decision trees, KNN, and Neural networks.

In [8] a comparative study of five ML models- Linear Regression, decision tree, random forest, Support Vector Machine (SVM) and least absolute shrinkage and selection operator (LASSO) is performed. Each model predicts the total active cases, total recoveries and total deaths in the imminent five days. The study shows that the use of these techniques is a promising strategy for forecasting. The best results are shown by poly linear regression and poly LASSO, followed by Linear regression, LASSO, Random forest and decision trees. SVM shows poor results in the study.

Another study by F. Rustam, et.al., shows a similar approach. They demonstrate that ML models are competent to forecast the number of upcoming COVID cases. Support vector machine (SVM), linear regression (LR), least absolute shrinkage and selection operator (LASSO), and exponential smoothing (ES) have been modeled[9]. Each model is used to predict the new infected cases, the no. of deaths, and the number of recoveries in the forthcoming 10 days. The results prove that the ES performs best among the modeled algorithms. LR and LASSO also perform considerably well whereas SVM is a poor performer for all three predictions.

While these algorithms give good results for short term for extended periods most of these algorithms have low reliability. In [10] the author to overcome this drawback proposes a deep learning approach. The author investigate the life cycle and spread of COVID-19 in Saudi Arabia and Bahrain using this approach. They found Support Vector regression (SVR) to be the best model for Saudi Arabia and linear regression works best for Bahrain. The LSTM model is also used to predict the cumulative reported, recovered and death cases globally.

C. Agent Based Model

An agent-based model simulates the proliferation of an infection through the social relations of the agents present within the environment. To design a realistic model world with similar environmental conditions and social culture the census data may be used. The agents are divided into different social groups. Each group has varying characteristics. Different types of family structures exist within the model world, that is, couples with and without kids, single parents, bachelors etc. These features are assigned such that it is proportional to the actual data. Social interactions are also modeled. For instance during the working hours students interact at schools and colleges, professionals interact with colleagues at work and retired population stays at home. Off working hours agents may come across other agents at house or neighborhood. During one of these encounters if an agent comes in contact to an infected agent there can be spread of disease.

At a given step the probability of an agent contracting the disease depends on contagiousness of the disease and number of sick agents interacted with in that step. Usually symptomatic cases are assumed to be more infectious. But this assumption may not hold good depending upon the disease, an example of which is Typhoid Mary. In case when the assumption holds good active period of the

disease is the period when agents infect the other agents they come in contact with. After the active period agents don't infect the other agents. Also symptoms are experienced by the agents after some time known as the incubation period of the disease. The start of epidemic is seeded randomly in agents near the airports assuming that the epidemic is started by agents entering the environment via international airports.

A classic example of an agent-based model is AceMod. It is an influenza spread model inspired by the 2009 swine flu pandemic. It is a generalized model. It models the social interactions of Australians to study the spread of infectious disease in such a environment.

Using AceMod as a precedent many have tried to model the COVID-19 Pandemic. One such effort is done by Maziarz et.al.,[7]. In this study the part specifying the social behavior of agents in Australia and the model society developed in 2016 for AceMod remain unchanged except for updating the population, whereas the assumption identifying infectivity of disease is changed. The attack rate, infectivity, reproductive number and various other parameters of this model are altered to better fit the features of corona virus.

The study claims that despite a comparison of the model predictions to the actual epidemic curve in Australia shows to be mismatched, this can be accounted for by taking into consideration that the government took strict steps very early on during the onset of the pandemic. Hence they suggest that the performance of an agent based model depends on the quality of the method used,i.e, how representative the model society is of the general populace, the implementation of the method,i.e., the programming, calibration and simulation of the model and the stability of the results,i.e., sensitivity of the results to changes in assumptions. While the first two criteria can be easily fulfilled the third criteria requires more data and would only be fulfilled in the later stages of an epidemic.

III. MODEL IMPLEMENTATION

Polynomial regression, Bayesian ridge regression and Support vector machine are the three models used in this paper. The steps followed in model implementation are:

1) Data Selection
2) Data Pre-processing and Feature Extraction
3) Model Building
4) Model validation

A. Dataset Used

COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[11]. This dataset is developed in response to the ongoing public health emergency to visualize and track reported cases in real-time. The dashboard was first shared publicly on January 22, illustrating the location and number of confirmed COVID-19 cases, deaths and recoveries for all affected countries. This repository is developed for research and tracking the pandemic. The latest data is available as GitHub repository and is updated daily.

B. Polynomial Regression

Polynomial regression is similar to linear regression that is it is statistically equivalent to multiple linear regressions. In linear regression the relation between independent and dependent variable is a polynomial of degree 1. In polynomial regression y is expressed by an nth degree polynomial in x. A 4^{th} degree polynomial is used in this model.

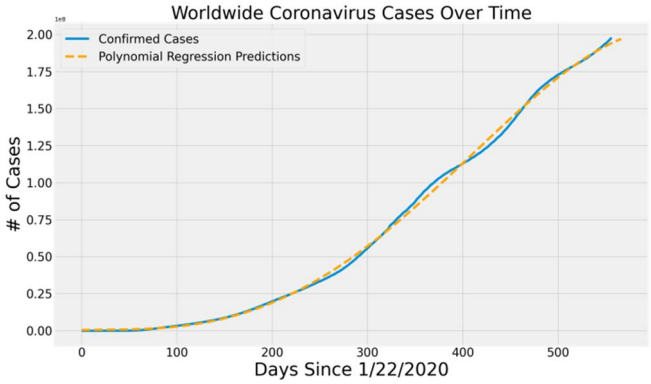


Figure 1: Predicted data vs. real data(PR)

Figure 1 shows the comparison of polynomial regression prediction with the real data. As can be seen the prediction of the model is very similar to the real data.

C. Bayesian Ridge Regression

Bayesian Ridge regression uses the Bayes theorem to find out the value of regression coefficients and the output is drawn from a probability distribution and not estimated as a single value. In this method of regression, the posterior distribution of the features is determined instead of finding the least-squares. In this algorithm prior domain knowledge can be incorporated to make the model more effective and if estimates are not present ahead of time, non-informative priors for the parameters can beused.

The results of the BLR are a distribution of possible model parameters based on the data and the prior. This allows quantification of uncertainty about the model, that is, if fewer data points are present then the posterior distribution will be more spread out. As the amount of data points increases, the likelihood washes out the prior, and in the case of infinite data, the outputs for the parameters converge to the values obtained from ordinary least squares. A 5^{th} degree Bayesian polynomial is used in this model.

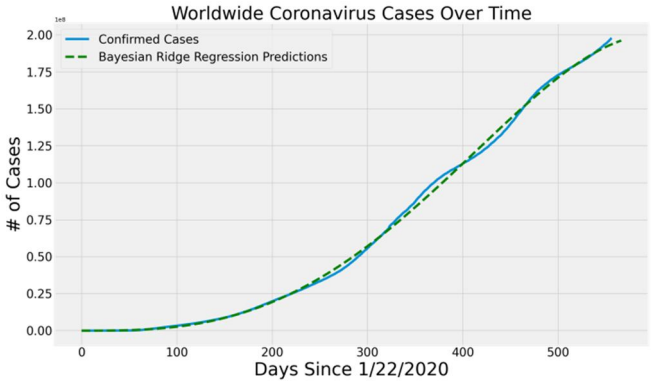


Figure 2: Predicted data vs. real data(BRR)

Figure 2 shows the comparison of Bayesian Ridge regression prediction with the real data. The prediction of the model has slight variations compared to the real data as can be seen.

D. Support Vector Machine

Support vector machine (SVM) is an algorithm that finds a hyperplane in N-dimensional space to classify the data points where N signifies the number features being used. The subset of points used by SVM for decision functions are called as support vectors. In geometrical approach, the dimension of hyperplane which is a subspace is one less than that of the classification space. This concept can be applied in any general space where the notion of dimension of subspace is defined.

There are various parameters used in SVM. For this model a poly kernel of degree 3 is used.

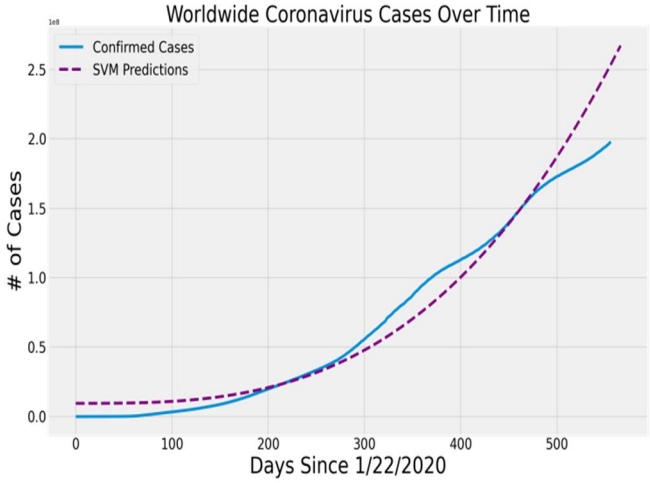


Figure 3: Predicted data vs. real data(SVM)

Figure 3 shows the comparison of SVM prediction with the real data. As can be seen the prediction of the model has large variations compared to the real data. Hence we can identify that SVM fails for the COVID-19 dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

IV. RESULTS AND DISCUSSION

The prediction of the three algorithms for the forthcoming 10 days from 31st July 2021 that is from 1st to 10th august is as shown in the Table 1.

DATE	Confirmed Cases prediction		
	Support vector machine	Polynomial regression	Bayesian ridge regression
01/08/2021	255139570.000000	194552401.000000	193966256.000000
02/08/2021	256464737.000000	194849032.000000	194235280.000000
03/08/2021	257794661.000000	195140929.000000	194498874.000000
04/08/2021	259129353.000000	195428049.000000	194756987.000000
05/08/2021	260468820.000000	195710352.000000	195009568.000000
06/08/2021	261813071.000000	195987796.000000	195256563.000000
07/08/2021	263162114.000000	196260339.000000	195497922.000000
08/08/2021	264515958.000000	196527940.000000	195733592.000000
09/08/2021	265874611.000000	196790556.000000	195963520.000000
10/08/2021	267238082.000000	197048145.000000	196187655.000000

Table 1: Prediction of confirmed cases from 1st to 10th august 2021

The performance metric of all algorithms with their respective performance values is as shown in Table 2.

Algorithm	Mean absolute error	Mean squared error
Polynomial Regression	1700762.1218629405	3970680911164.969
Bayesian ridge Regression	2085357.1649777554	5645192051980.205
Support vector machine	50616020.745460075	2573104215758295.5

Table 2: Performance Metric of the Models

To understand the performance of the model better we utilize the Mean absolute error (MAE) and Mean square error (MSE) as the performance metric. It can be identified that polynomial regression has the best performance using both MAE and MSE.It is followed by Bayesian Ridge Regression. SVM has a poor performance.

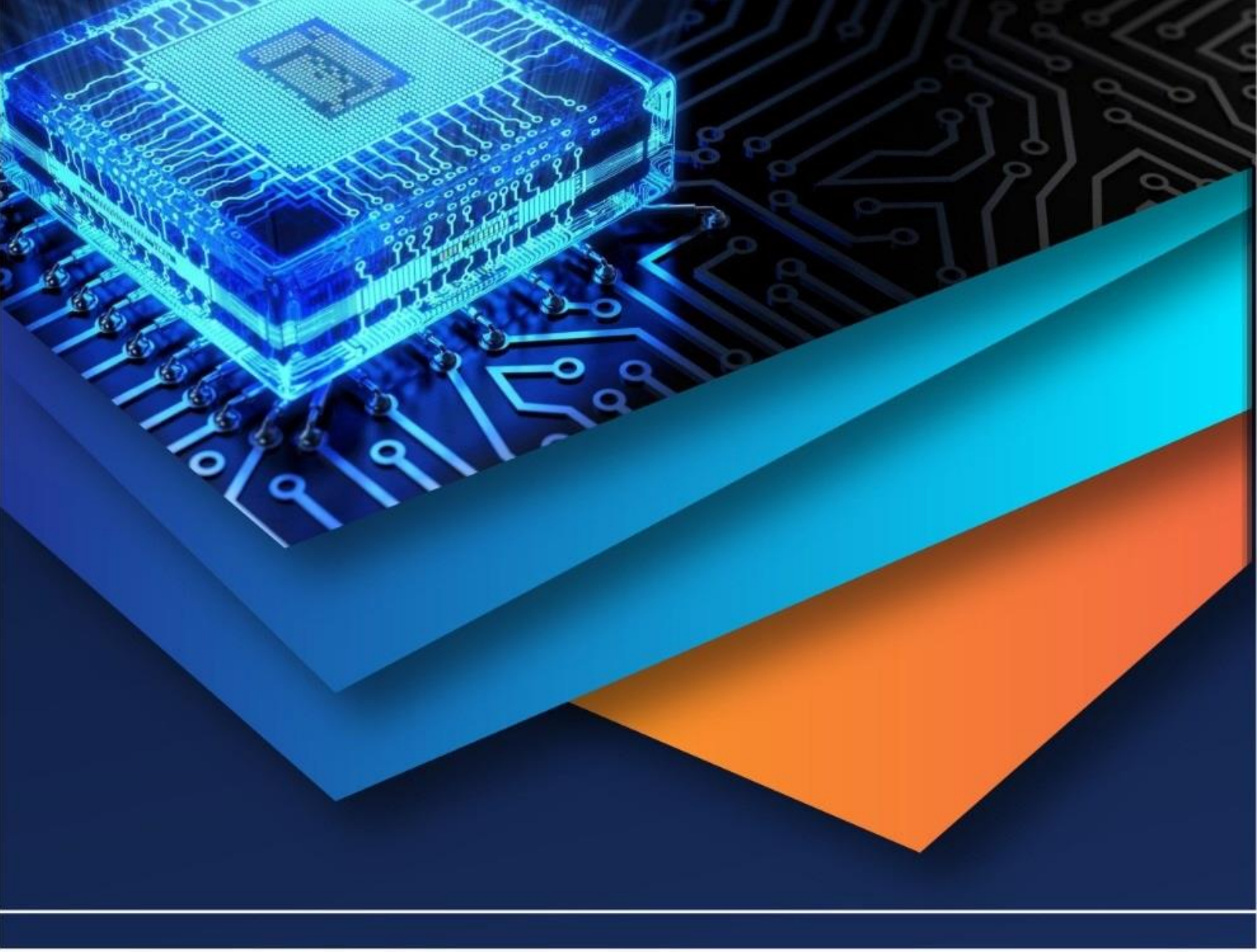
V. CONCLUSION AND FUTURE SCOPE

This project simulates the COVID-19 epidemic trend using the conventional machine learning models using the available data. It is found that Polynomial Regression has the best performance followed by Bayesian ridge regression. Support Vector Machine has a poor performance. This study demonstrates that forecasting the trend of an epidemic is possible using Machine Learning. This can be utilized by the government to be better prepared with resources to minimize the death toll. This study can also help in taking appropriate steps to reduce the spread as much as possible by studying the data of other countries and analyzing the steps taken by countries that see decrease in COVID cases. The idea of this project is to study the connection of data of many days to predict a pattern. The same can be utilized by health officers and government to make provisions for future. However the number of COVID cases is heavily influenced by many unquantifiable factors such as environmental factors, government restrictions, awareness among masses etc. Hence there is a need to introduce these to the modelling to gain better insights for the future. This model is a generalized model, but for better results models can be made taking local factors into account for smaller populations. It may yield better results as COVID has different patterns in different regions influenced by climate, population density and behavior of local population.

Also there have been different variants of corona virus that have been in circulation. This Factor can also be incorporated in the study as each of these variants has different contagiousness and the severity also varies. Hence, more research is required in this field with lot of scope for exploration.

REFERENCES

[1] Jo H, Son H, Hwang HJ, Jung SY. Analysis of COVID-19 spread in South Korea using the SIR model with time-dependent parameters and deep learning. medRxiv; 2020. DOI: 10.1101/2020.04.13.20063412
[2] https://www.medrxiv.org/content/10.1101/2020.06.02.20119917v2.full.pdf
[3] Acemoglu D, Chernozhukov V, Werning I, Whinston MD (2020) A multi-risk SIR model with optimally targeted lockdown. NBER working paper 27102
[4] Hethcote, H.; Zhien, M.; Shengbing, L. Effects of quarantine in six endemic models for infectious diseases. Mathematical biosciences
[5] Islam MS, Irana Ira J, Ariful Kabir KM, Kamrujjaman M. 2020. COVID-19 epidemic compartments model and Bangladesh. Preprint. (2020).
[6] A. Tiwari, Modelling and analysis of covid-19 epidemic in india.,medRxiv (2020).
[7] Maziarz, Mariusz, and Martin Zach. "Agent-based modelling for SARS-CoV-2 epidemic prediction and intervention assessment: A methodological appraisal." Journal of evaluation in clinical practice vol. 26,5 (2020): 1352-1360. doi:10.1111/jep.13459
[8] Vartika Bhadana, et al. "A Comparative Study of Machine Learning Models for COVID-19 Prediction in India." 2020 IEEE 4th Conference on Information & Communication Technology (CICT),2020. Crossref, doi:10.1109/cict51604.2020.9312112.
[9] F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311
[10] H. Khaloofi, J. Hussain, Z. Azhar and H. F. Ahmad, "Performance Evaluation of Machine Learning Approaches for COVID-19 Forecasting by Infectious Disease Modeling," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430192.
[11] https://github.com/CSSEGISandData/COVID-19



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY